

# In search of lost introns

Miklós Csűrös\*, J. Andrew Holey†, Igor B. Rogozin‡

February 6, 2008

## Abstract

Many fundamental questions concerning the emergence and subsequent evolution of eukaryotic exon-intron organization are still unsettled. Genome-scale comparative studies, which can shed light on crucial aspects of eukaryotic evolution, require adequate computational tools.

We describe novel computational methods for studying spliceosomal intron evolution. Our goal is to give a reliable characterization of the dynamics of intron evolution. Our algorithmic innovations address the identification of orthologous introns, and the likelihood-based analysis of intron data. We discuss a compression method for the evaluation of the likelihood function, which is noteworthy for phylogenetic likelihood problems in general. We prove that after  $O(n\ell)$  preprocessing time, subsequent evaluations take  $O(n\ell/\log \ell)$  time almost surely in the Yule-Harding random model of  $n$ -taxon phylogenies, where  $\ell$  is the input sequence length.

We illustrate the practicality of our methods by compiling and analyzing a data set involving 18 eukaryotes, more than in any other study to date. The study yields the surprising result that ancestral eukaryotes were fairly intron-rich. For example, the bilaterian ancestor is estimated to have had more than 90% as many introns as vertebrates do now.

---

\*Department of Computer Science and Operations Research, Université de Montréal, Québec, Canada

†Department of Computer Science, Saint John's University and the College of St. Benedict, Collegeville, Minn., USA

‡National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Md., USA

**Contact:** csuros AT iro.umontreal.ca

## 1 Introduction

Typical eukaryotic protein-coding genes contain introns, which are removed prior to translation. Key constituents of the spliceosome, which is the RNA-protein complex that performs the intron excision, can be traced back (Collins and Penny 2005) to the last common ancestor of extant eukaryotes (LECA). Even deep-branching lineages (Vaňáčová et al. 2005; Nixon et al. 2002) have introns. It is thus almost certain that spliceosomal introns were present in LECA. Moreover, when comparing distant eukaryotes, intron positions often agree (Rogozin et al. 2003). The similarity is likely due more to conservation of early introns than to parallel gains (Roy and Gilbert 2005; Sverdlov et al. 2005). It is thus compelling to use genome-scale comparisons to study intron evolution in different lineages, and even to estimate the exon-intron organization in extinct ancestors. One of the first such studies, by Rogozin et al. (2003), involved orthologous gene sets in eight eukaryotes. The same data set was reanalyzed by different authors (Roy and Gilbert 2005; Csűrös 2005; Carmel et al. 2005; Nguyen et al. 2005), using novel methods developed for intron data. Subsequent inquiries (Nielsen et al. 2004; Roy and Penny 2006; Roy and Penny 2007; Coulombe-Huntington and Majewski 2007) attest to a renewed interest in understanding the specifics of intron evolution within different eukaryotic lineages. This paper introduces novel computational techniques for the analysis of spliceosomal intron evolution, anticipating more large-scale studies to come.

Section 2 describes an alignment method for intron-annotated protein sequences, as well as a segmentation method for identifying conserved portions of a multiple alignment. Section 3 describes a likelihood framework in which intron evolution can be analyzed in a theoretically sound manner. Section 4 scrutinizes a compression technique that accelerates the evaluation of the likelihood function. The compression involves an  $O(n\ell)$ -time preprocessing step for  $\ell$  sites and a phylogeny with  $n$  species. We show that the subsequent evaluation takes sublinear,  $O(n\ell/\log \ell)$  time almost surely in the Yule-Harding model of random phylogenies, even in the case of arbitrary, constant-size alphabets. Fast evaluation is particularly important when the likelihood is maximized in a numerical procedure that computes the likelihood function with many different parameter settings. Section 5 describes two applications

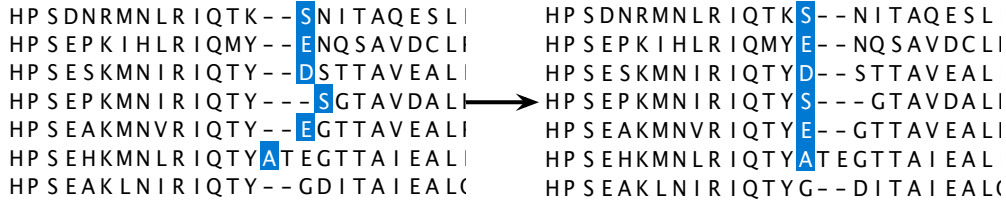


Figure 1: Fragment of a multiple alignment before and after realignment using intron annotation. Shaded rectangles show the intron positions projected to the protein sequences.

of our methods. In one application, intron-aware alignment was used to validate some unexpected features of intron evolution before LECA. In a second application, we analyzed intron evolution in 18 eukaryotic species. We found evidence of intron-rich early eukaryotes and a prevalence of intron loss over intron gain in recent times.

## 2 Identification of orthologous introns

### 2.1 Intron-aware alignment

Orthologous introns can be identified by using whole genome alignments in the case of closely related genomes (Coulombe-Huntington and Majewski 2007). For distant eukaryotes, however, intron orthology can only be established through protein alignments (Rogozin et al. 2005). The usual procedure is to project intron positions onto an alignment of multiple orthologous proteins. If introns in different species are projected to the same alignment position, then the introns are assumed to be orthologous.

Intron annotation can be included in protein alignments by defining intron match and mismatch scores. The alignment score is then computed as the sum of scores for amino acid matches, gaps, and intron matches. Incorporating intron annotation should lead to better alignments at the amino acid level, and to a more reliable identification of orthologous introns. Figure 1 shows an example of such improvement.

Intron scoring can be based on log-likelihood ratios in a probabilistic model (Durbin et al. 1998). The model is defined by the joint distribution  $p$  for the intron state in two sequences  $S$  and  $T$ , and the prior distributions  $\pi_S$

and  $\pi_T$ . Aligned sites have states  $(s, t)$  with probability  $\pi_S(s) \cdot \pi_T(t)$  if the two sites are unrelated, or with probability  $p(s, t)$  in case of homology. An  $(s, t)$  alignment is scored with a value that is proportional to  $-\log \frac{p(s, t)}{\pi_S(s)\pi_T(t)}$ .

We used the data set of Rogozin et al. (2003) to assess the strength of intron-match signals. Since the data include no sites in which all species lack introns, but the model does allow for that, we added extra sites with no introns. The original data comprise 7236 intron sites in 684 genes, across 8 species. Using a method described earlier (Csürös 2005), we added 35000 unobserved intron sites. Using estimates for  $p$  and  $\pi$ , we computed the appropriate scores. The intron score is asymmetric and varies with evolutionary distance and intron conservation. Matches for absent introns have an insignificant score, but shared introns have a high score, such as 93 (human-Plasmodium), 106 (human-Arabidopsis), 152 (human-*S. pombe*) or 303 (Drosophila-Anopheles) on a 1/60-bit scale. Shared introns thus give a signal comparable to amino acid matches: in the 1/60-bit scaled version of the VTML240 matrix (Müller et al. 2002), a tryptophan match scores 289, and an arginine identity scores 113.

Consider the case of aligning two protein sequences,  $S$  and  $T$ , which are annotated with the intron positions. Every residue has two associated intron sites (after the first and second nucleotides of their codons), and there is an intron site between consecutive amino acid positions (phase-0 introns). Intron sites may or may not be filled in by introns in either sequence. We use the notation for  $S[i: 0]$  for the phase-0 site preceding the codon for amino acid  $i$ , and  $S[i: 1]$ ,  $S[i: 2]$  for phase-1 and -2 sites within the codon. Intron presence is encoded by 1, and intron absence is encoded by 0 throughout the paper. The intron annotation is specified by the variables  $S[i: j] \in \{1, 0\}$ . (There can be no introns after the last amino acid.) Scores for aligned introns are specified by a  $2 \times 2$  scoring matrix  $\Lambda$ .

Phase-1 and phase-2 intron sites are automatically placed by their associated amino acids. If  $M$  is the amino acid scoring matrix, then the alignment of  $S[i]$  and  $T[j]$  entails a score of  $M \begin{bmatrix} S[i] \\ T[j] \end{bmatrix} + \Lambda \begin{bmatrix} S[i] \\ T[j] \end{bmatrix}$  with  $\Lambda \begin{bmatrix} S[i] \\ T[j] \end{bmatrix} = \Lambda \begin{bmatrix} S[i: 1] \\ T[j: 1] \end{bmatrix} + \Lambda \begin{bmatrix} S[i: 2] \\ T[j: 2] \end{bmatrix}$ . Similarly, aligning  $S[i]$  with an indel implies a score of  $M \begin{bmatrix} S[i] \\ - \end{bmatrix} + \Lambda \begin{bmatrix} S[i] \\ 0 \end{bmatrix} = M \begin{bmatrix} S[i] \\ - \end{bmatrix} + \Lambda \begin{bmatrix} S[i: 1] \\ 0 \end{bmatrix} + \Lambda \begin{bmatrix} S[i: 2] \\ 0 \end{bmatrix}$ , in addition to possible gap opening and closing penalties. Standard alignment procedures need to be modified to deal with phase-0 introns, since the placement of phase-0

introns is not fixed with respect to gaps. It is not possible to simply add a new character to the alphabet to represent phase-0 introns because they affect gaps differently from amino acids.

We added intron scoring into a multiple alignment framework, using a sum-of-pairs scoring policy with affine gap-scoring. It is NP-hard to find the alignment of two multiple alignments under these optimization criteria (Ma et al. 2003). Even the alignment of a single sequence to a multiple alignment necessitates sophisticated techniques (Kececioğlu and Zhang 1998). Our solution therefore uses a gap-counting heuristic: namely, a gap-open penalty is triggered for an indel aligned with an amino acid if the indel is preceded by an amino acid or a phase-0 intron. Gap opening thus corresponds to a pattern  $\begin{smallmatrix} \text{x} \\ 1- \end{smallmatrix}$  or  $\begin{smallmatrix} *? \text{x} \\ \text{x}0- \end{smallmatrix}$ . Here, 1, 0 are intron states for the phase-0 site, and ? denotes either state. In addition, x denotes an arbitrary amino acid, - is the indel character, and \* is either of the latter two. We implemented affine gap-scoring by separate gap-open and -close penalties, so that gaps at the alignment extremities can be penalized less severely. Gap closing is counted for the patterns  $\begin{smallmatrix} -1 \\ \text{x}? \end{smallmatrix}$  and  $\begin{smallmatrix} -0 \text{x} \\ \text{x}?* \end{smallmatrix}$ .

Table 1 gives the recurrences for a dynamic programming algorithm that aligns an intron-annotated sequence  $S$  to a multiple alignment  $P$  of  $h$  intron-annotated protein sequences. In order to simplify the presentation, we represent the sequences in such a way that every odd position of  $S$  and  $P$  is a regular residue or alignment column, annotated with information on the presence of phase-1 and phase-2 introns, whereas every even position is a phase-0 intron site. We use  $\Lambda \begin{bmatrix} S[i] \\ P[j] \end{bmatrix}$  to denote the sum of intron-match scores for the intron sites associated with the positions  $S[i]$ ,  $P[j]$ . We use also the shorthand  $M \begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}$  to denote scoring for the alignment of an amino acid  $x$  with an amino acid profile  $\mathbf{y}$ . The algorithm uses three types of variables,  $A[i, j]$ ,  $\mathbf{gS}[i, j]$  and  $\mathbf{gP}[i, j]$ , which correspond to partial prefix alignments ending with aligned residues, gaps in  $S$ , or gaps in  $P$ , respectively. In case of  $\mathbf{gS}[i, j]$ , the last indel must be aligned with an amino acid column, and, thus  $j$  must be odd; for  $\mathbf{gP}[i, j]$ ,  $i$  must be odd.

Gaps are scored by using affine penalties, with gap-open, -extend, and -close scores, denoted by  $\gamma^{(<)}, \gamma^{(-)}, \gamma^{(>)}$ . The gap-counting heuristic implies that gap scores in the equations of Table 1 are defined by the number of certain patterns in up to three consecutive alignment columns. For instance,  $\gamma_2^{(<)}(j)$  equals the gap-open penalty multiplied by the number of such rows

$$\begin{aligned}
A[i, j] &= M \left[ \begin{smallmatrix} S[i] \\ P[j] \end{smallmatrix} \right] + \Lambda \left[ \begin{smallmatrix} S[i] \\ P[j] \end{smallmatrix} \right] + \gamma_1^{(-)}(j) + \max \left\{ A[i-2, j-2] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ P[j-1] \end{smallmatrix} \right] + \gamma_1^{(<)}(j) + \gamma_1^{(>)}(j), \right. \\
&\quad \text{gS}[i-2, j-2] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ P[j-1] \end{smallmatrix} \right] + \gamma_1^{(<)}(j) + \gamma_2^{(>)}(j-2), \text{gS}[i-1, j-2] + \Lambda \left[ \begin{smallmatrix} 0 \\ P[j-1] \end{smallmatrix} \right] + \gamma_1^{(<)}(j) + \gamma_2^{(>)}(j-2), \\
&\quad \left. \text{gP}[i-2, j-2] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ P[j-1] \end{smallmatrix} \right] + \gamma_2^{(<)}(j) + \gamma_3^{(>)}(j), \text{gP}[i-2, j-1] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ 0^h \end{smallmatrix} \right] + \gamma_2^{(>)}(j) \right\} \\
\text{gS}[i, j] &= \Lambda \left[ \begin{smallmatrix} 0 \\ P[j] \end{smallmatrix} \right] + \gamma_2^{(-)}(j) + \max \left\{ A[i, j-2] + \Lambda \left[ \begin{smallmatrix} 0 \\ P[j-1] \end{smallmatrix} \right] + \gamma_3^{(<)}(j) + \gamma_1^{(>)}(j), \right. \\
&\quad \text{gS}[i, j-2] + \Lambda \left[ \begin{smallmatrix} 0 \\ P[j-1] \end{smallmatrix} \right], \\
&\quad \left. \text{gP}[i, j-2] + \Lambda \left[ \begin{smallmatrix} 0 \\ P[j-1] \end{smallmatrix} \right] + \gamma_3^{(<)}(j) + \gamma_3^{(>)}(j), \text{gP}[i, j-1] + \Lambda \left[ \begin{smallmatrix} 0 \\ 0^h \end{smallmatrix} \right] + \gamma_3^{(<)}(j) + \gamma_2^{(>)}(j) \right\} \\
\text{gP}[i, j] &= \Lambda \left[ \begin{smallmatrix} S[i] \\ 0^h \end{smallmatrix} \right] + \gamma_1^{(-)}(j) + \max \left\{ A[i-2, j] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ 0^h \end{smallmatrix} \right] + \gamma_3^{(<)}(j), \right. \\
&\quad \text{gS}[i-2, j] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ 0^h \end{smallmatrix} \right] + \gamma_3^{(<)}(j) + \gamma_2^{(>)}(j), \text{gS}[i-1, j] + \Lambda \left[ \begin{smallmatrix} 0 \\ 0^h \end{smallmatrix} \right] + \gamma_3^{(<)}(j) + \gamma_2^{(>)}(j), \\
&\quad \left. \text{gP}[i-2, j] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ 0^h \end{smallmatrix} \right] \right\} \\
\text{gS}[i, j] &= \Lambda \left[ \begin{smallmatrix} 0 \\ P[j] \end{smallmatrix} \right] + \gamma_2^{(-)}(j) + \max \left\{ A[i-1, j-2] + \Lambda \left[ \begin{smallmatrix} S[i] \\ P[j-1] \end{smallmatrix} \right] + \gamma_3^{(<)}(j) + \gamma_1^{(>)}(j), \right. \\
&\quad \text{gS}[i-1, j-2] + \Lambda \left[ \begin{smallmatrix} S[i] \\ P[j-1] \end{smallmatrix} \right] + \gamma_4^{(<)}(i, j) + \gamma_4^{(>)}(i, j), \text{gS}[i, j-2] + \Lambda \left[ \begin{smallmatrix} 0 \\ P[j-1] \end{smallmatrix} \right], \\
&\quad \left. \text{gP}[i-1, j-2] + \Lambda \left[ \begin{smallmatrix} S[i] \\ P[j-1] \end{smallmatrix} \right] + \gamma_3^{(<)}(j) + \gamma_3^{(>)}(j), \text{gP}[i-1, j-1] + \Lambda \left[ \begin{smallmatrix} S[i] \\ 0^h \end{smallmatrix} \right] + \gamma_3^{(<)}(j) + \gamma_2^{(>)}(j) \right\} \\
\text{gP}[i, j] &= \Lambda \left[ \begin{smallmatrix} S[i] \\ 0^h \end{smallmatrix} \right] + \gamma_1^{(-)}(j) + \max \left\{ A[i-2, j-1] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ P[j] \end{smallmatrix} \right] + \gamma_5^{(<)}(j) + \gamma_5^{(>)}(j), \right. \\
&\quad \text{gS}[i-2, j-1] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ P[j] \end{smallmatrix} \right] + \gamma_5^{(<)}(j) + \gamma_2^{(>)}(j-1), \text{gS}[i-1, j-1] + \Lambda \left[ \begin{smallmatrix} 0 \\ P[j] \end{smallmatrix} \right] + \gamma_5^{(<)}(j) + \gamma_2^{(>)}(j-1), \\
&\quad \left. \text{gP}[i-2, j-1] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ P[j] \end{smallmatrix} \right] + \gamma_6^{(<)}(j) + \gamma_6^{(>)}(j), \text{gP}[i-2, j] + \Lambda \left[ \begin{smallmatrix} S[i-1] \\ 0^h \end{smallmatrix} \right] \right\}
\end{aligned}$$

Table 1: Recurrences for intron-aware alignment. Odd positions correspond to regular amino acids and possibly phase-1 and -2 introns. Even positions are placeholders for phase-0 intron sites.

$\gamma_1^{(<)}$	1- or x0-	$\gamma_2^{(<)}$	1-
$\gamma_3^{(<)}$	x	$\gamma_4^{(<)}$	x if $S[i]$ has intron, otherwise nothing
$\gamma_5^{(<)}$	1 or x0	$\gamma_6^{(<)}$	1
$\gamma_1^{(>)}$	-0x or -1*	$\gamma_2^{(>)}$	x
$\gamma_3^{(>)}$	1* or 0x	$\gamma_4^{(>)}$	x if $S[i]$ has intron, otherwise nothing
$\gamma_5^{(>)}$	-1	$\gamma_6^{(>)}$	1
$\gamma_1^{(-)}$	-	$\gamma_2^{(-)}$	x

Table 2: Patterns for gap counting. The index  $j$  in  $\gamma^{(<)}(j)$ ,  $\gamma^{(-)}(j)$  and  $\gamma^{(>)}(j)$  is the index for the last column in the pattern.

in  $P$  where column  $j$  contains an indel, and column  $j-1$  has a phase-0 intron. The corresponding pattern is described as 1-. Table 2 lists the patterns for the gap-counting heuristic.

Table 1 does not show the initialization of the variables, nor the final gap-counting: they employ a logic analogous to the recurrences. At the end of the algorithm, the best of  $A[|S|, |P|]$ ,  $gS[|S|, |P|]$ ,  $gP[|S|, |P|]$  is selected, and the actual alignment is found by standard traceback techniques (Durbin et al. 1998).

We implemented the algorithm in Java. The program iteratively realigns one sequence at a time to the rest of the sequences in a multiple alignment. Instead of sequence-dependent intron match-mismatch scores, the implementation uses only two parameters: one for intron conservation and another for intron loss/gain.

## 2.2 Identification of conserved blocks

In order to reliably identify orthologs, we need to be able to distinguish regions of the alignment that are highly conserved from those that are less well-conserved. In poorly conserved regions, we cannot confidently infer intron orthology.

Zhang et al. (1999) proposed post-processing pairwise sequence alignments into alternating blocks that score above a threshold parameter  $\alpha$  or below  $(-\alpha)$ . We attained a similar goal by adapting algorithmic techniques from Cs  r  s (2004). The procedure separates a multiple alignment into alternating high- and low-scoring regions. Using a complexity penalty  $\alpha$ , a

segmentation with  $k$  high-scoring regions has a segmentation score of  $A - k\alpha$  where  $A$  is the sum of scores for the aligned columns. Column scores are computed without gap-open and -close penalties. The best segmentation of an alignment of length  $\ell$  can be found in  $O(\ell)$  time after the column scores are computed.

The result of this computation is that the total of column scores in each selected high-scoring region will be greater than  $\alpha$ . There may be small sub-regions of negative scores, but the total score of such a sub-region cannot be less than  $(-\alpha)$ . Conversely, unselected regions score below  $(-\alpha)$  and cannot have sub-regions scoring above  $\alpha$ .

### 3 A likelihood framework

#### 3.1 Markov models of evolution

We use a Markov model for intron evolution, as in previous studies (Csűrös 2005). For the sake of generality, we describe the Markov model (Steel 1994; Felsenstein 2004) over an arbitrary alphabet  $\mathcal{A}$  of fixed size  $r = |\mathcal{A}|$ . The intron alphabet is  $\mathcal{A} = \{0, 1\}$ . A *phylogeny* over a set of species  $X$  is defined by a rooted tree  $T$  and a probabilistic model. The leaves are bijectively mapped to elements of  $X$ . Each tree node  $u$  has an associated random variable  $\xi(u)$ , which is called its *state* or *label*, that takes values over  $\mathcal{A}$ . The tree  $T$  with its parameters defines the joint distribution for the random variables  $\xi(u)$ . The distribution is determined by the *root probabilities*  $(\pi(a) : a \in \mathcal{A})$  and *edge transition probabilities*  $(p_e(a \rightarrow b) : a, b \in \mathcal{A})$  assigned to every edge  $e$ . The root probabilities give the distribution of the root state. Edge transition probabilities define the conditional distributions  $\mathbb{P}\left\{\xi(u_{i+1}) = b \mid \xi(u_i) = a\right\} = p_{u_i u_{i+1}}(a \rightarrow b)$ . Along every path away from the root, the node states form a Markov chain. The leaf states form the *character*  $\xi = (\xi(u) : u \in X)$ . An input data set (or *sample*) consists of independent and identically distributed (iid) characters:  $\mathbf{\xi} = (\xi_i : i = 1, \dots, \ell)$ .

In case of intron evolution, introns are generated by a two-state continuous-time Markov process with *gain* and *loss rates*  $\lambda_e, \mu_e \geq 0$  along each edge  $e$ . The edge length is denoted by  $t_e$ . Using standard results (Ross 1996), the transition probabilities on edge  $e$  with rates  $\lambda_e = \lambda, \mu_e = \mu$  and length  $t_e = t$  can be written as  $p_e(0 \rightarrow 1) = \frac{\lambda}{\lambda + \mu}(1 - e^{-t(\lambda + \mu)})$ ,  $p_e(1 \rightarrow 0) = \frac{\mu}{\lambda + \mu}(1 -$



$e^{-t(\lambda+\mu)}$ ), and  $p_e(0 \rightarrow 0) = 1 - p_e(0 \rightarrow 1)$ ,  $p_e(1 \rightarrow 1) = 1 - p_e(1 \rightarrow 0)$ . In the absence of established edge lengths, we fix the edge length scaling in such a way that  $\lambda_e + \mu_e = 1$ . Independent model parameters are thus  $\pi(1)$ ,  $\nu_e$  and  $t_e$  for all edges  $e$ . It is important to allow for branch-dependent rates, since loss and gain rates vary considerably between lineages (Jeffares et al. 2006; Roy and Gilbert 2006).

In a maximum likelihood approach, model parameters are set by maximizing the likelihood of the input sample. Let  $\mathbf{x} = (x_1, \dots, x_\ell)$  be the input data. Every  $x_i$  is a vector of  $n$  states, and we write  $x_i(u)$  to denote the observed state of leaf  $u$ . By independence, the likelihood of  $\mathbf{x}$  is the product  $\mathbb{P}\{\boldsymbol{\xi} = \mathbf{x}\} = \prod_i \mathbb{P}\{\xi = x_i\}$ . Each character's likelihood can be computed in  $O(n)$  time, using a dynamic programming procedure introduced by Felsenstein (1981). The procedure relies on a “pruning” technique, which consists of computing the conditional likelihoods  $L_u(a)$  for every node  $u$  and letter  $a$ .  $L_u(a)$  is the probability of observing the leaf labelings in the subtree of  $u$ , given that  $u$  is labeled with  $a$ . The likelihood for the character  $x$  equals  $L(x) = \mathbb{P}\{\xi = x\} = \sum_{a \in \mathcal{A}} \pi(a) L_{\text{root}}(a)$ .

Intron data are somewhat unusual in that an all-0 character is never observed: the input does not include sites in which introns are absent in all of the organisms. The uncorrected likelihood function is therefore misleading, as it underestimates the probability of intron loss. To resolve this difficulty, we employ a correction technique proposed by Felsenstein (1992) for restriction sites. (Csűrös (2005) describes an alternative technique based on expectation maximization.) We compute the likelihood under the condition that the input does not include all-0 characters. We use therefore the corrected likelihood  $L'(x) = \frac{L(x)}{1 - L(0^n)}$ , and maximize  $L' = \prod_i L'(x_i)$ .

## 3.2 Ancestral events in intron evolution

Our goal is to give a reliable characterization of the dynamics of intron evolution. In particular, we aim to give estimates for intron density in ancestral species, and for intron loss and gain events on the edges. Notice that the estimation method needs to account for ancestral introns even if they got eliminated in all descendant lineages. It is possible to do that with the help of conditional expectations, which fit naturally into a likelihood framework.

For an observed character  $x$ , we define *upper conditional likelihoods*  $U_u(a)$  so that  $U_u(a)L_u(a) = \mathbb{P}\{\xi = x, \xi(u) = a\}$ . Upper conditional likelihoods are computed with dynamic programming, from the root towards the leaves,

in  $O(n\ell)$  time (Csürös 2005), even for non-binary trees. Similar computations are routinely used in DNA and protein likelihood maximization programs (Adachi and Hasegawa 1995; Guindon and Gascuel 2003). Here we allow irreversible probabilistic models, which explains why  $U_u(a)$  must be computed in a top-down fashion in (1).

The posterior probability for the state at node  $u$  is

$$q_a^{(u)}(x) = \mathbb{P}\left\{\xi(u) = a \mid \xi = x\right\} = \frac{U_u(a)L_u(a)}{\sum_{b \in \mathcal{A}} U_u(b)L_u(b)}.$$

The posterior probabilities for state changes on an edge  $uv$  are

$$\begin{aligned} q_{ab}^{(v)}(x) &= \mathbb{P}\left\{\xi(u) = a, \xi(v) = b \mid \xi = x\right\} \\ &= U_u(a)L_u(a) \frac{p_{uv}(a \rightarrow b)L_v(b)}{\sum_{a'} p_{uv}(a \rightarrow a')L_v(a')}. \end{aligned}$$

Now, the number of ancestral introns is estimated as the conditional expectation  $N_u = \ell_0 q_1^{(u)}(0^n) + \sum_i q_1^{(u)}(x_i)$ . The formula takes into consideration unobserved intron sites, by estimating their number  $\ell_0 = \ell \frac{L(0^n)}{1-L(0^n)}$  as the mean of a negative binomial random variable. The number of intron state changes is estimated as  $N_v(a \rightarrow b) = \ell_0 q_{ab}^{(v)}(0^n) + \sum_i q_{ab}^{(v)}(x_i)$ . In particular,  $N_v(1 \rightarrow 0)$  is the number of introns lost, and  $N_v(0 \rightarrow 1)$  is the number of introns gained on the edge leading to  $v$ .

In order to compute  $U$ , we initialize  $U_{\text{root}}(a) = \pi(a)$ . On every edge  $uv$ ,

$$\begin{aligned} U_v(b) &= \sum_{a \in \mathcal{A}} U_u(a) p_{uv}(a \rightarrow b) \prod_{w \in \text{siblings}(v)} \sum_{a' \in \mathcal{A}} p_{uw}(a \rightarrow a') L_w(a') \\ &= \sum_{a \in \mathcal{A}} U_u(a) L_u(a) \frac{p_{uv}(a \rightarrow b) L_v(b)}{\sum_{a' \in \mathcal{A}} p_{uv}(a \rightarrow a') L_v(a')}. \quad (1) \end{aligned}$$

## 4 Rapid computation of the likelihood

There are many heuristics that accelerate likelihood-based phylogenetic reconstruction (Friedman et al. 2002; Guindon and Gascuel 2003), which mostly concentrate on the exploration of the tree space. We propose an improvement to the evaluation of the likelihood function, which normally takes linear

time in the input size. Our evaluation method yields an  $O(\log \ell)$  speedup for typical trees. We use it to optimize the parameters of intron evolution on a known tree, but the method is generally applicable to phylogenetic likelihood problems where the likelihood is numerically optimized. The key idea is that it is enough to carry out the pruning algorithm once for every different labeling within a subtree. Different labelings within each subtree can be computed in a preprocessing step. Subsequent evaluations of the likelihood function with different model parameters are faster, and depend on the number of different labelings in the data set.

Here we describe how the preprocessing step can be carried out in  $O(n\ell)$  time. Secondly, we analyze the computational complexity of subsequent evaluations, and show that an  $O(\log \ell)$  speedup is achieved for almost all random trees in the Yule-Harding model. The latter analysis produces some novel concentration results on the number of subtrees with a fixed size  $k$ .

To our knowledge, the closest idea to ours was articulated by Stamatakis et al. (2002). Specifically, they proposed identifying characters in which leaves in a subtree have identical labels. They reported that in benchmark experiments with nucleotide sequences, likelihood optimization was accelerated by 12–15% through this technique. The technique relies entirely on the fact that alignments of closely related sequences exhibit high levels of identity, and cannot be extended to non-identical labelings.

## 4.1 Yule-Harding model

The Yule-Harding distribution is encountered in random birth and death models of species and in coalescent models (Felsenstein 2004), and is thus one of the most adequate random models for phylogenies. In one of the equivalent formulations, a random tree is grown by adding leaves one by one in a random order. The leaves are first numbered by using a random uniform permutation of the integers  $1, 2, \dots, n$ . Leaves are joined to the tree in an iterative procedure. In step 1, the tree is just leaf 1 on its own. In step 2, the tree is a “cherry” with leaves 1 and 2. In each subsequent step  $i = 3, \dots, n$ , a random leaf  $Y_i$  is picked uniformly from the set  $\{1, 2, \dots, i - 1\}$ . The new leaf  $i$  is added to the tree as the sibling of  $Y_i$ , forming a cherry: a new node is placed on the edge leading to  $Y_i$  and  $i$  is connected to it. The resulting random tree in iteration  $n$  has the Yule-Harding distribution (Harding 1971).

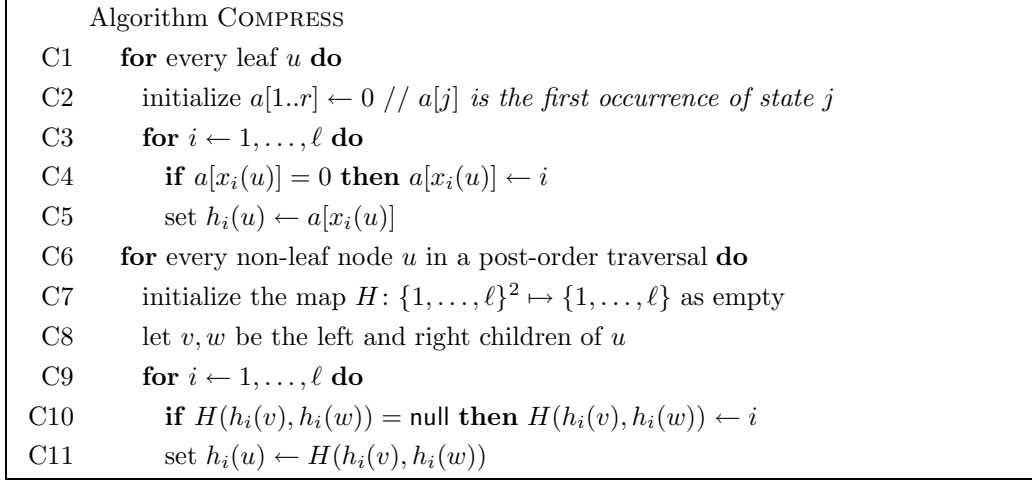


Figure 2: Computing the auxiliary arrays from which the multiplicities  $\nu_u$  are obtained.

## 4.2 Preprocessing

The likelihood can be computed faster by first identifying the different  $x_i$  values, along with their multiplicity in the data. For large trees, the input typically consists of many different labelings, but for small trees with  $n < \log_r \ell$ , the compression is useful, since the number of different  $x_i$  values is bounded by  $r^n < \ell$ . In order to exploit the benefits of compression, we extend it to every subtree.

**Definition 1.** Define the multiset of observed labelings for every node  $u$  as follows. Let  $n'$  denote the number of leaves in the subtree rooted at  $u$ , and let  $u_1, \dots, u_{n'}$  denote those leaves. Define  $\nu_u(y)$  for every labeling  $y \in \mathcal{A}^{n'}$  of the leaves in the subtree of  $u$  as the number of times  $y$  is observed in the data:

$$\nu_u(y) = \sum_{i=1}^{\ell} \chi \left\{ \forall k: y_k = x_i(u_k) \right\},$$

where  $\chi\{\cdot\}$  denotes the indicator function that takes the value 1 if its argument is true, otherwise it is 0. Define also the set of observed labelings

$$\mathcal{S}_u = \left\{ y \in \mathcal{A}^{n'} : \nu_u(y) > 0 \right\}.$$

The multisets of observed labelings are computed in the preprocessing step. The likelihood function is evaluated subsequently by computing the

Algorithm LOGLIKELIHOOD	
L1	<b>for</b> every leaf $u$ <b>do</b>
L2	set $L_u[a][a'] \leftarrow \chi\{a = a'\}$ for all $a \in \mathcal{S}_u$ and $a' \in \mathcal{A}$
L3	<b>for</b> every non-leaf node $u$ in a post-order traversal <b>do</b>
L4	<b>for</b> every labeling $y \in \mathcal{S}_u$ <b>do</b>
L5	let $u_1, u_2$ be the children of $u$ , and let $y_1, y_2$ be their subtree labelings
L6	<b>for</b> every $a \in \mathcal{A}$
L7	$L_u[y][a] \leftarrow \prod_{j=1,2} \sum_{a' \in \mathcal{A}} p_{uu_j}(a \rightarrow a') L_{u_j}[y_j][a']$
L8	set $\text{logL} \leftarrow 0$
L9	<b>for</b> every labeling $y \in \mathcal{S}_{\text{root}}$ <b>do</b>
L10	$\text{logL} \leftarrow \text{logL} + \nu_{\text{root}}(y) \cdot \sum_{a \in \mathcal{A}} \log(\pi(a) L_{\text{root}}[y][a])$
L11	<b>return</b> $\text{logL}$

Figure 3: Computing the log-likelihood using the observed labelings.

conditional likelihoods at each node  $u$  for the labelings of  $\mathcal{S}_u$  only, in  $O(|\mathcal{S}_u|)$  time.

In order to compute  $\nu_u$ , we use a recursive procedure. It is important to avoid working with the  $O(n)$ -dimensional vectors  $y$  of Definition 1 directly, otherwise the preprocessing may take superlinear time in  $n$ . For that reason, every labeling  $y \in \mathcal{S}_u$  is represented by the index  $i$  for which  $x_i$  is the first occurrence of  $y$ . Accordingly, we compute an auxiliary array  $h_i(u)$ , which stores the first occurrence of each labeling  $x_i$  in  $u$ 's subtree. In particular,  $h_i(u) = i'$  if  $i'$  is the smallest index such that  $x_i(u_k) = x_{i'}(u_k)$  for all  $k$  where  $u_k$  are the leaves in  $u$ 's subtree. Figure 2 shows that the values  $h_i(u)$  can be computed in a post-order traversal. After  $h_i(u)$  are computed for all  $i$  and  $u$ , the multiplicities  $\nu_u$  and observed labelings  $\mathcal{S}_u$  are straightforward to calculate in linear time. The map  $H$  in Lines C7–C11 is sparse with at most  $\ell$  entries, and can be implemented as a hash table so that accessing and updating it takes  $O(1)$  time. Consequently, Algorithm COMPRESS takes  $O(n\ell)$  time.

### 4.3 Evaluating the likelihood function

After the preprocessing step, the conditional likelihoods are computed only for the different labelings within each subtree. Figure 3 shows the evaluation of the likelihood function. The running time for the algorithm is  $O(s)$  where  $s$

is the total number of different labelings within all subtrees:

$$s = \sum_u |\mathcal{S}_u|.$$

If  $n_u$  denotes the number of leaves in the subtree rooted at  $u$ , then  $\mathcal{S}_u$  has at most  $r^{n_u}$  elements. Hence,  $s$  is bounded as

$$s \leq \sum_u \min\{r^{n_u}, \ell\}. \quad (2)$$

Observe that by sheer number of arithmetic operations, it is always worth evaluating the likelihood function this way. The worst situation is that of a caterpillar tree (where every inner node has a leaf child). In that case, there are only a few ( $\lfloor \log_r \ell \rfloor - 1$ ) non-leaf nodes for which we can compress the data, and it is possible to construct an artificial data set in which there are  $\ell$  different labelings for  $n - O(\log \ell)$  subtrees. Caterpillar trees, however, are rare in phylogenetic analysis. Typical phylogenies have fairly balanced subtrees (Aldous 2001).

In what follows, we examine the bound of (2) more closely in the Yule-Harding model. The analysis relies on a characterization of the random number of subtrees with a given size, as expressed in Theorems 1 and 2.

**Theorem 1.** *Consider random evolutionary trees with  $n$  leaves in the Yule-Harding model. Let  $C_k$  denote the number of subtrees with  $k = 1, \dots, n-1$  leaves in a random tree. The expected value of  $C_k$  is*

$$\mathbb{E}C_k = 2n \left( \frac{1}{k} - \frac{1}{k+1} \right). \quad (3)$$

Trivially,  $\mathbb{E}C_n = 1$ .

*Proof.* Heard (1992) derives Equation (3) by appealing to a Pólya urn model. An equivalent result is stated by Devroye (1991). □

**Theorem 2.** *For all  $\epsilon > 0$ ,*

$$\mathbb{P}\{C_k \leq \mathbb{E}C_k - \epsilon\} \leq e^{-\frac{\epsilon^2}{2n}}; \quad (4a)$$

$$\mathbb{P}\{C_k \geq \mathbb{E}C_k + \epsilon\} \leq e^{-\frac{\epsilon^2}{2n}}. \quad (4b)$$

*Proof.* Consider the random construction of the tree. Let  $Y_i$  denote the random leaf picked in step  $i$  to which leaf  $i$  gets connected, for  $i = 3, 4, \dots, n$ . Each random variable  $Y_i$  is uniformly distributed over the set  $\{1, 2, \dots, i-1\}$ , and  $Y_3, Y_4, \dots, Y_n$  are independent. Moreover, they completely determine the tree  $T$  at the end of the procedure. Consequently,  $C_k$  is a function of  $(Y_i: i = 3, \dots, n)$ :  $C_k = f(Y_3, Y_4, \dots, Y_n)$ . The key observation for the concentration result is that if we change the value of only one of the  $Y_i$  in the series, then  $C_k$  changes by at most two. In order to see this, consider what happens to the tree  $T$ , if we change the value of exactly one of the  $Y_i$  from  $y$  to  $y'$ . Such a change corresponds to a “subtree prune and regraft” transformation (Felsenstein 2004). Specifically, the subtree  $T_i$ , defined as the child tree of the lowest common ancestor  $u$  of  $y$  and  $i$  containing the leaf  $i$ , is cut from  $T$ , and is reattached to one of the edges on the path from the root to  $y'$ . Now, such a transformation does not affect  $C_k$  by much. Notice that subtree sizes are strictly monotone decreasing from the root on every path. On the path from the root to  $u$ , subtree sizes decrease by the size  $\tau$  of  $T_i$ , and  $u$  disappears, contributing a change of  $+1$ ,  $0$  or  $(-1)$  to  $C_k$ . (At most one subtree of size  $\tau + k$  that contains  $T_i$  now has size  $k$ , and at most one subtree of original size  $k$  is not counted anymore: it may be  $u$ ’s subtree itself, or a subtree above it.) An analogous argument shows that regrafting contributes a change of  $+1$ ,  $0$ , or  $(-1)$ . Hence, the function  $f(\cdot)$  is such that by changing one of its arguments, its value changes by at most 2. As a consequence, McDiarmid’s inequality ((1989)) can be applied to bound the probabilities of large deviations for  $C_k$ . In particular, for all  $\epsilon > 0$ ,

$$\mathbb{P}\left\{f(Y_3, \dots, Y_n) - \mathbb{E}f(Y_3, \dots, Y_n) \leq -\epsilon\right\} \leq e^{-2\epsilon^2/c^2}$$

where  $c^2 = \sum_{i=3}^n c_i^2$  and

$$c_i = \max_{y_3, \dots, y_n, y, y'} \left| f(y_3, \dots, y_{i-1}, y, y_{i+1}, \dots, y_n) - f(y_3, \dots, y_{i-1}, y', y_{i+1}, \dots, y_n) \right|.$$

Since  $c_i \leq 2$  for all  $i$ ,  $\mathbb{P}\left\{C_k \leq \mathbb{E}C_k - \epsilon\right\} \leq e^{-\frac{\epsilon^2}{2(n-2)}}$ , implying Eq. (4a). An identical bound holds for the right-hand tail of the distribution.  $\square \quad \square$

REMARK. The particular case of  $k = 2$  was considered by McKenzie and Steel (2000). They showed that the distribution of  $C_2$  is asymptotically normal with mean  $n/3$

$n$	$\ell$	$r$	$n\ell$	$n \mathcal{S}_{\text{root}} $	bound	$s$
8	7236	2	101304	1386	368	183
18	8044	2	273496	19142	16764	1196
47	5216	3	479872	309120	65743	10305

Table 3: Effect of the compression on three data sets. The fourth column quantifies the direct evaluation method, the fifth column quantifies the effect of the compression restricted to the root, the sixth column corresponds to the bound of Eq. (2) and the seventh column gives the exact value of  $s$ . The first data set is from Rogozin et al. (2003), the second data set is the one analyzed here in §5.2, and the third one is an unpublished data set we have worked on, where an ambiguity character is included in the alphabet.

and variance  $2n/45$ . The result suggests that the best constant factor in the exponent of Eqs. (4) is  $45/8$  for  $i = 2$ , instead of  $\frac{1}{2}$ . Rosenberg (2006) gave exact formulas for the variance of  $C_k$ . He showed that the variance of  $C_k$  is  $(2 + o(1))\frac{n}{k^2}$ . The variance formulas were given earlier in a different context by Devroye (1991). (See also the discussion by Blum and François (2005).) The result suggests that by analogy with the cherries, the best constant factor in the exponent is  $(1/4 + o(1))k^2$ . It is thus plausible that the probability is properly bounded by  $1 - o(n \log^{-2} \ell)$  in Theorem 3 below.

**Theorem 3.** *With probability  $1 - o\left(\frac{n}{\log^4 \ell}\right)$ , the likelihood function can be evaluated for random trees in the Yule-Harding model in  $O\left(\frac{n\ell}{\log \ell}\right)$  time after an initial preprocessing step that takes  $O(n\ell)$  time. Evaluating the likelihood function takes  $O(n\ell)$  time in the worst case, and  $O(n\ell \log^{-1} \ell)$  time on average.*

We need the following lemma for the proof of Theorem 3.

**Lemma 4.** *For all  $t \geq 4$ ,  $\sum_{k=1}^t \frac{2^k}{k(k+1)} < \frac{2^t}{t+1}$ . For all  $t \geq 1$  and  $r = 3, 4, \dots$ ,  $\sum_{k=1}^t \frac{r^k}{k(k+1)} \leq \frac{r^t}{t+1}$ .*

*Proof.* The proof is straightforward by induction in  $t$ . Notice that the right order of magnitude is  $\sum_{k=1}^t \frac{r^k}{k(k+1)} = \Theta\left(t^{-2}r^t\right)$  but we need a bound for all  $t$ .  $\square$   $\square$



*Proof of Theorem 3.* The preprocessing (Fig. 2) takes  $O(n\ell)$  time as discussed in §4.2. The evaluation of the likelihood function (Fig. 3) takes  $O(s)$  time. By Eq. (2),

$$s \leq \sum_{k=1}^n C_k \min\{r^k, \ell\} = \sum_{k=1}^{\lfloor \log_r \ell \rfloor} C_k r^k + \sum_{k=1+\lfloor \log_r \ell \rfloor}^n C_k \ell. \quad (5)$$

Let  $t = \lfloor \log_r \ell \rfloor$ . By Theorem 1 and Lemma 4, if  $\ell \geq 16$  or  $r \geq 3$ ,

$$\begin{aligned} \mathbb{E}s &\leq 2n \sum_{k=1}^t \frac{r^k}{k(k+1)} + \ell \left( \frac{2n}{t+1} - 1 \right) \\ &\leq 2n \frac{r^t}{t+1} + 2n \frac{\ell}{t+1} \leq \frac{4n\ell}{t+1}, \end{aligned} \quad (6)$$

which proves our claim about the average running time. Now, let  $\epsilon = \frac{n}{2t(t+1)}$ . Plugging  $\epsilon$  into Theorem 2, we get that

$$\mathbb{P}\left\{|C_k - \mathbb{E}C_k| \geq \epsilon\right\} \leq 2 \exp\left(-\frac{n}{8} \left(\frac{1}{t} - \frac{1}{t+1}\right)^2\right). \quad (7)$$

Let  $\mathcal{E}_k$  denote the event that  $|C_k - \mathbb{E}C_k| < \epsilon$  for  $k = 1, \dots, t$ , and let  $\mathcal{E}_{t+1}$  denote the event that  $|\sum_{k=1+t}^n C_k - \mathbb{E} \sum_{k=1+t}^n C_k| < t\epsilon$ . Since  $\sum_k C_k = 2n - 1$ ,  $\cap_{k=1}^t \mathcal{E}_k$  implies  $\mathcal{E}_{t+1}$ . By (7),

$$\mathbb{P} \bigcap_{k=1}^{t+1} \mathcal{E}_k \geq 1 - \sum_{k=1}^t \mathbb{P} \bar{\mathcal{E}}_k \geq 1 - 2t \exp\left(-\frac{n}{8} \left(\frac{1}{t} - \frac{1}{t+1}\right)^2\right),$$

where  $\bar{\mathcal{E}}_k$  denotes the complementary event to  $\mathcal{E}_k$ . Now,  $\cap_{k=1}^t \mathcal{E}_k$  also implies that  $s \leq \frac{5n\ell}{t+1}$ . Since the likelihood computation takes  $O(s)$  time, the theorem holds.  $\square$

Theorem 3 underestimates the actual speedups that the compression method brings about. Table 3 shows that compression results in a 50–500 fold speedup of the likelihood evaluation in practice. Notice also that the constants hidden behind the asymptotic notation are quite different between the preprocessing and evaluation steps: costly floating point operations are avoided in the preprocessing step. It is important to stress that the theoretical analysis does not rely on similarities in the input data: Theorem 3 and

the bound of (2) hold for any sample of size  $\ell$ . Real-life data are expected to behave even better, as Table 3 illustrates.

Even though the theorem holds for arbitrary alphabets, the compression is less effective for large alphabets (amino acids for example) in practice. For DNA sequences, however, it should still be valuable: we conjecture that compression would accelerate likelihood optimization by at least an order of magnitude.

We implemented Algorithms COMPRESS and LOGLIKELIHOOD, along with likelihood maximization and the posterior calculations of §3.2 in a Java package. Likelihood is maximized by setting loss and gain parameters for the edges, by using mostly the Broyden-Fletcher-Goldfarb-Shanno method (Press et al. 1997) whenever possible, and occasionally Brent’s line minimization (Press et al. 1997) for each parameter separately.

## 5 Applications

### 5.1 Ancient paralogs

In our first example, intron-aware alignment was used to reject a hypothesis about whether lack of intron sharing between homologous genes is due to poor protein alignments.

We used intron-aware alignment in a study about ancient eukaryotic paralogs (Sverdlov et al. 2007). In the study, 157 homologous gene families were examined across six species (*A. thaliana*, *H. sapiens*, *C. elegans*, *D. melanogaster*, *S. cerevisæ* and *S. pombe*). These families were notable because they contained paralogous members in multiple eukaryotic species, but not in prokaryotes, and, thus, presumably underwent duplication in the lineage leading to LECA. Ancient paralogs within and across species share very few (in the order of a few percentages) introns. The finding is quite surprising, as recent paralogs, resulting from lineage-specific duplications, and also orthologs between human and Arabidopsis agree much more in their intron sites (Rogozin et al. 2003). Consequently, ancient paralogs either lacked introns at the time of their duplication, or their duplication involved removal of introns.

In one of the data validation steps for the study, we used intron-aware alignment with very high intron match rewards. With larger rewards for intron matches, more introns line up in the alignment, but the protein align-

ment gets worse. Even by corrupting the protein alignment, we were not able to achieve intron sharing levels similar to that of human-Arabidopsis orthologs. The lack of intron agreement is therefore not an artifact of the protein alignments.

More details of our study will be described in a forthcoming publication (Sverdlov et al. 2007).

## 5.2 Intron-rich ancestors

We compiled a data set with 18 eukaryotic species to give a comprehensive picture of spliceosomal evolution among Eukaryotes.

Species	Abbreviation	Assembly	Source
<i>Homo sapiens</i>	Hsap	36.2	R
<i>Rattus norvegicus</i>	Rnor	RGSC 3.4	E
<i>Takifugu rubripes</i>	Trub	FUGU 4.0	E
<i>Danio rerio</i>	Drer	Zv6	E
<i>Drosophila melanogaster</i>	Dmel	BDGP 4.3	R
<i>Anopheles gambiae</i>	Agam	AgamP3	R
<i>Apis mellifera</i>	Amel	AMEL4.0	R
<i>Cænorhabditis elegans</i>	Cele	WS160	R
<i>Cænorhabditis briggsæ</i>	Cbri	CB25	W
<i>Saccharomyces cerevisiæ</i>	Scer	2.1	R
<i>Neurospora crassa</i> OR74A	Ncer		R
<i>Schizosaccharomyces pombe</i> 972h-	Spom	1.1	R
<i>Ustilago maydis</i> 521	Umay		R
<i>Cryptococcus neoformans</i> v. n. JEC21	Cneo	1.1	R
<i>Oryza sativa</i> ssp. <i>japonica</i>	Osat	RAP 3	R
<i>Arabidopsis thaliana</i>	Atha	6.0	R
<i>Plasmodium falciparum</i> 3D7	Pfal	1.1	R
<i>Plasmodium berghei</i> str. ANKA	Pber		R

Table 4: Data sources and species abbreviations. R: RefSeq release 20, E: Ensemble release 42, W: WormBase release 160.

### Data preparation

Genbank flatfiles and protein sequences were downloaded from RefSeq (Pruitt et al. 2007) and Ensembl (Hubbard et al. 2007). Exon-intron annotation was extracted from the Genbank flatfiles. The *C. briggsæ* protein sequences and genome annotation were downloaded from WormBase (Bieri et al. 2007), and intron

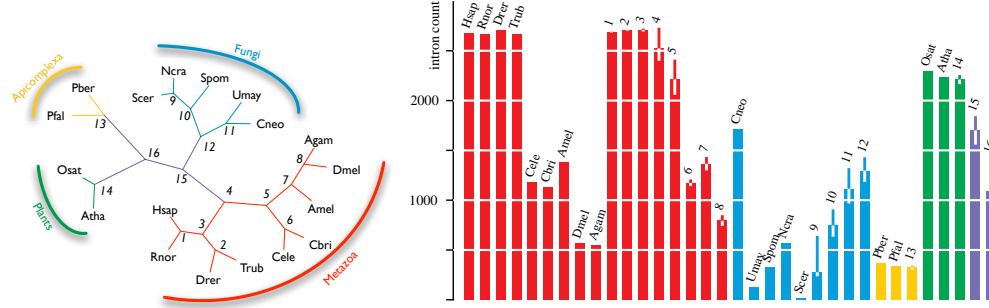


Figure 4: Phylogeny for the 18 species and estimated intron counts at ancestors. Ancestral nodes are identified by the numbers 1–16. Error bars denote 95% confidence intervals computed from 1000 simulated data sets. We rooted the tree at node 16 (Bikonts) for computational purposes.

annotation was extracted from the GFF annotation file. Table 4 lists the data sources and the species abbreviations. We used the 684 ortholog sets, each corresponding to a cluster of orthologous groups, or KOG (Tatusov et al. 2003), from the study of Rogozin et al. (2003) as “seeds” for compiling a set of putative orthologs for our species. For each seed (consisting of homologous protein sequences for eight species), we performed a position-specific iterated BLAST search (Altschul et al. 1997). In case of *Plasmodium* species, we used three iterations against a database of all protozoan peptide sequences in RefSeq. We used an E-value cutoff of  $10^{-9}$  for retaining candidates. Each candidate hit was then used as a query in a reversed position-specific BLAST (rpsblast) search (Marchler-Bauer et al. 2007) against the KOG database. Candidates were retained after this point if they had the highest scoring hit (by rpsblast) against the same KOG as the KOG of the seed data, and they scored within 80% of the best such hit for the species.

From the resulting set of paralogs, we selected a putative ortholog set in the following manner. For each KOG, we selected all possible triples of human-Arabidopsis-Saccharomyces paralogs and kept the triple that had the highest score in alignments built by MUSCLE (Edgar 2004). Alignment score was computed using the VTML240 matrix (Müller et al. 2002). Additional putative orthologs were added for one species at a time, by aligning each paralog individually to the current profile, and keeping the one that gave the

largest alignment score. At this iterative addition, scoring was done with the VTML240 matrix, by summing the five highest pairwise scores between a candidate and already included sequences.

The resulting sets were then realigned using MUSCLE, and then realigned again using our intron-aware alignment with a gap penalty of 300, gap-extend penalty of 11, VTML240 amino acid scoring, intron-match scores of 300 and intron-mismatch penalties of 20. (These latter were established using different intron-match and -mismatch scores, and selecting the ones that gave the fewest number of intron sites while decreasing the score of the implied protein alignment by less than 0.1%.)

Conserved portions were extracted using our segmentation program with a complexity penalty of  $\alpha = 400$  (larger values gave identical segmentation results, and lower values resulted in too many scattered blocks). We penalized indels with an infinitely large value to exclude gap columns. Phase-0 introns falling on the boundaries of conserved blocks were excluded. Intron presence and absence in the aligned data was then extracted to produce the data for the likelihood programs.

## Results

Figure 4 shows the estimated intron densities for ancestral species. It is notable that ancestral nodes such as the bilaterian ancestor (node 4), the ecdysozoan ancestor (node 5), the opisthokont ancestor (node 15), and the bikont ancestor (node 16) all have very high intron densities, surpassing most previous estimates (Rogozin et al. 2003; Csűrös 2005). The ecdysozoan ancestor has an even higher estimated intron density (80% of human density) than the otherwise quite generous estimates (about 70%) of Roy and Gilbert (2005), which is mostly due to the inclusion of the relatively intron-rich honeybee genes (IHBSC 2006). Intron density in the bilaterian ancestor is estimated to be almost as high as in humans (94%), agreeing with estimates of Roy and Gilbert (2005). Sequences of a handful of intron-rich genes from the marine annelid *Platynereis dumerilii* have already indicated that the bilaterian ancestor’s genome was at least two-thirds as intron-rich as the vertebrate genomes even by conservative estimates (Raible et al. 2005; Roy 2006).

Introns are for the most part lost on the branches. Figure 5 shows the estimated changes in a few lineages. Intron evolution has been much slower in the vertebrate lineage than in most other lineages: in more than 380 million

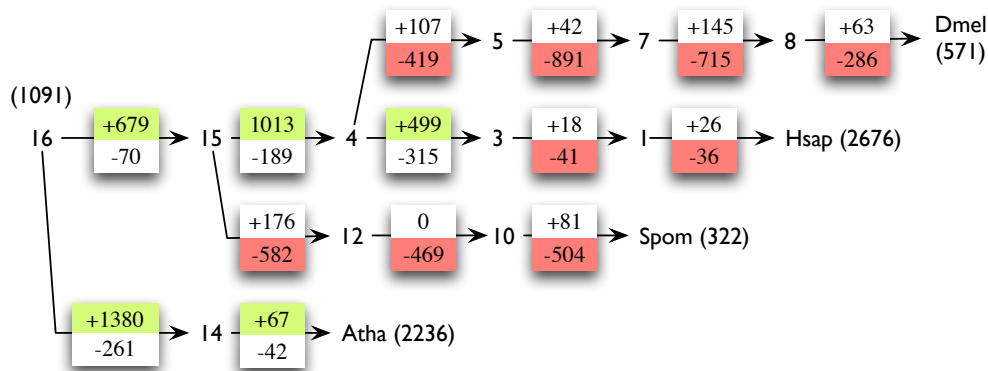


Figure 5: Intron gains and losses in a few evolutionary paths. Estimated and actual intron counts are in parentheses.

years since the divergence with fishes, only about 3% of our introns got lost. Fungi, for example, massively trimmed their introns in many lineages. A notable exception is *C. neoformans*, which seems to have gained introns, but that assessment may change if another basidiomycete genome becomes available besides the relatively intron-poor *U. maydis*.

## 6 Conclusion

We presented a novel alignment technique for establishing intron orthology, and a likelihood framework in which intron evolutionary events can be quantified. We described a compression method for the evaluation of the likelihood function, which has been extremely valuable in practice. We also showed that the compression leads to sublinear running times for likelihood evaluation.

We illustrated our methods for analyzing intron evolution with a large and diverse set of eukaryotic organisms. The data set is more comprehensive than any used in other studies published to this day. The data indicate that ancestral eukaryotic genomes were more intron-rich than previous studies suggested.

Many circumstances influence intron loss (Jeffares et al. 2006; Roy and Gilbert 2006), and realistic likelihood models need to introduce rate variation (Carmel et al. 2005). Usual rate variation models (Felsenstein 2004) entail multiple evaluations of the likelihood function, and, thus, underline the importance of computational

efficiency. We believe that the proposed methods will help to produce and analyze large data sets even within complicated likelihood models.

## Acknowledgment

This research is supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

## References

- Adachi, J. and M. Hasegawa (1995). MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Volume 28 of *Computer Science Monographs*, pp. 1–150. Tokyo, Japan: Institute of Statistical Mathematics.
- Aldous, D. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science* 16, 23–34.
- Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17), 3389–3402.
- Bieri, T. et al. (2007). WormBase: new content and better access. *Nucleic Acids Research* 35(suppl\_1), D506–510.
- Blum, M. G. B. and O. François (2005). On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. *Mathematical Biosciences* 195, 141–153.
- Carmel, L., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin (2005). An expectation-maximization algorithm for analysis of evolution of exon-intron structure of eukaryotic genes. See McLysaght and Huson (2005), pp. 35–46.
- Collins, L. and D. Penny (2005). Complex spliceosomal organization ancestral to extant eukaryotes. *Molecular Biology and Evolution* 22(4), 1053–1066.
- Coulombe-Huntington, J. and J. Majewski (2007). Characterization of intron loss events in mammals. *Genome Research* 17(1), 23–32.

- Csűrös, M. (2004). Maximum-scoring segment sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1(4), 139–150.
- Csűrös, M. (2005). Likely scenarios of intron evolution. See McLysaght and Huson (2005), pp. 47–60. DOI:10.1007/11554714\_5.
- Devroye, L. (1991). Limit laws for local counters in random binary search trees. *Random Structures and Algorithms* 2(1), 303–315.
- Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. UK: Cambridge University Press.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5), 1792–1797.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17, 368–376.
- Felsenstein, J. (1992). Phylogenies from restriction sites, a maximum likelihood approach. *Evolution* 46, 159–173.
- Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, Mass.: Sinauer Associates.
- Friedman, N., M. Ninio, T. Pupko, I. Pe’er, and T. Pupko (2002). A structural EM algorithm for phylogenetic inference. *Journal of Computational Biology* 9(2), 331–353.
- Guindon, S. and O. Gascuel (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52(5), 696–704.
- Harding, E. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability* 3, 44–77.
- Heard, S. B. (1992). Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution* 46(6), 1818–1826.
- Hubbard, T. J. P. et al. (2007). Ensembl 2007. *Nucleic Acids Research* 35(suppl\_1), D610–617.
- IHBSC (2006). Insights into social insects from the genome of the honey bee *Apis mellifera*. *Nature* 443, 931–949.



- Jeffares, D. C., T. Mourier, and D. Penny (2006). The biology of intron gain and loss. *Trends in Genetics* 22(1), 16–22.
- Kececioğlu, J. and W. Zhang (1998). Aligning alignments. In M. Farach-Colton (Ed.), *Proc. CPM*, Volume 1448 of *LNCS*, pp. 189–208. Springer.
- Ma, B., Z. Wang, and K. Zhang (2003). Alignment between two multiple alignments. In R. A. Baeza-Yates, E. Chávez, and M. Crochemore (Eds.), *Proc. Combinatorial Pattern Matching (CPM)*, Volume 2676 of *LNCS*, pp. 254–265. Springer.
- Marchler-Bauer, A. et al. (2007). CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Research* 35(suppl\_1), D237–240.
- McDiarmid, C. (1989). On the method of bounded differences. In J. Siemons (Ed.), *Surveys in Combinatorics*, pp. 148–184. Cambridge University Press.
- McKenzie, A. and M. Steel (2000). Distributions of cherries for two models of trees. *Mathematical Biosciences* 164, 81–92.
- McLysaght, A. and D. Huson (Eds.) (2005). *Proc. RECOMB Satellite Workshop on Comparative Genomics*, Volume 3678 of *LNCS*. Springer-Verlag.
- Müller, T., R. Spang, and M. Vingron (2002). Estimating amino acid substitution models: a comparison of Dayhoff’s estimator, the resolvent approach and a maximum likelihood method. *Molecular Biology and Evolution* 19(1), 8–13.
- Nguyen, H. D., M. Yoshihama, and N. Kenmochi (2005). New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Computational Biology* 1(7), e79.
- Nielsen, C. B., B. Friendman, B. Birren, C. B. Burge, and J. E. Galagan (2004). Patterns of intron gain and loss in fungi. *PLoS Biology* 2(12), e422.
- Nixon, J. E. J., A. Wang, H. G. Morrison, A. G. McArthur, M. L. Sogin, B. J. Loftus, and J. Samuelson (2002). A spliceosomal intron in *Giardia lamblia*. *Proceedings of the National Academy of Sciences of the USA* 99(6), 3701–3705.

- Press, W. H., S. A. Teukolsky, W. V. Vetterling, and B. P. Flannery (1997). *Numerical Recipes in C: The Art of Scientific Computing* (Second ed.). Cambridge University Press.
- Pruitt, K. D., T. Tatusova, and D. R. Maglott (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35(suppl\_1), D61–65.
- Raible, F., K. Tessmar-Raible, K. Osoegawa, P. Wincker, C. Jubin, G. Balavoine, D. Ferrier, V. Benes, P. de Jong, J. Weissenbach, P. Bork, and D. Arendt (2005). Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science* 310, 1325–1326.
- Rogozin, I. B., A. V. Sverdlov, V. N. Babenko, and E. V. Koonin (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. *Briefings in Bioinformatics* 6(2), 118–134.
- Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology* 13, 1512–1517.
- Rosenberg, N. A. (2006). The mean and variance of  $r$ -pronged nodes and  $r$ -caterpillars in Yule-generated genealogies. *Annals of Combinatorics* 10, 129–146.
- Ross, S. M. (1996). *Stochastic Processes* (Second ed.). Wiley & Sons.
- Roy, S. W. (2006). Intron-rich ancestors. *Trends in Genetics* 22(9), 468–471.
- Roy, S. W. and W. Gilbert (2005). Complex early genes. *Proceedings of the National Academy of Sciences of the USA* 102(6), 1986–1991.
- Roy, S. W. and W. Gilbert (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* 7, 211–221.
- Roy, S. W. and D. Penny (2006). Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution. *Genome Research* 16(10), 1270–1275.
- Roy, S. W. and D. Penny (2007). Patterns of intron loss and gain in plants: Intron loss-dominated evolution and genome-wide comparison

- of *O. sativa* and *A. thaliana*. *Molecular Biology and Evolution* 24(1), 171–181.
- Stamatakis, A. P., T. Ludwig, H. Meier, and M. J. Wolf (2002). AxML: A fast program for sequential and parallel phylogenetic tree calculations based on the maximum likelihood method. In *Proc. IEEE Computer Society Bioinformatics Conference (CSB)*, pp. 21–28.
- Steel, M. A. (1994). Recovering a tree from the leaf colourations it generates under a Markov model. *Applied Mathematics Letters* 7, 19–24.
- Sverdlov, A. V., M. Csűrös, I. B. Rogozin, and E. V. Koonin (2007). A glimpse of a putative pre-intron phase of eukaryotic evolution. *Trends in Genetics*. In press. DOI: /10.1016/j.tig.2007.01.001.
- Sverdlov, A. V., I. B. Rogozin, V. N. Babenko, and E. V. Koonin (2005). Conservation versus parallel gains in intron evolution. *Nucleic Acids Research* 33(6), 1741–1748.
- Tatusov, R. L. et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 441.
- Vaňáčová, Š., W. Yan, J. M. Carlton, and P. J. Johnson (2005). Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. *Proceedings of the National Academy of Sciences of the USA* 102(12), 4430–4435.
- Zhang, Z., P. Berman, T. Wiehe, and W. Miller (1999). Post-processing long pairwise alignments. *Bioinformatics* 15(12), 1012–1019.